Final Review and Practice Problems

Shannon Pileggi

STAT 217

OUTLINE

Final exam

The final exam *is* cumulative

- $\sim$ 20% Unit 1
- $\sim$ 20% Unit 2
- $\sim$ 60% Unit 3

Please bring your calculator! Questions?

Topics covered

- Descriptive Statistics ◂ More
- Study Design ◂ More
- Associations ◂ More
- Distributions and Probability ◂ More
- Sampling Distributions ◂ More
- Confidence Intervals ◂ confidence intervals
- Hypothesis Tests ◂ steps ◂ errors
- Different Methods ◂ overview ◂ tests ◂ calculations

Which of the following are true statements about p-values? Mark all that apply.

1. A nonsignificant difference (eg, p-value>0.05) means that the null hypothesis is true.
2. Sample size can affect your p-value.
3. A scientific conclusion should be based solely on whether or not the p-value is significant.

Final Remarks

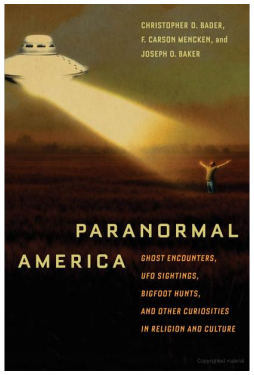
- quantitative methods/statistics are applied in *all* disciplines, regardless of whether you are in the humanities, social sciences, or natural sciences
- a p-value isn't the end of the story
  - association does not mean causation
  - a statistically significant result isn't always meaningful
- sometimes statistical analysis is the end result of the research, but sometimes it is just the beginning....

- established in 1991
- largest conference on UFOs in US
- registration ~$200

- >20 speakers discussing topics related to the UFO phenomenon including technology, government cover-ups, exopolitics, black projects, crop circles, alien visitation and more
- speakers include astrophysicists, nuclear physicists, abductees, and former top-secret-clearance military personnel

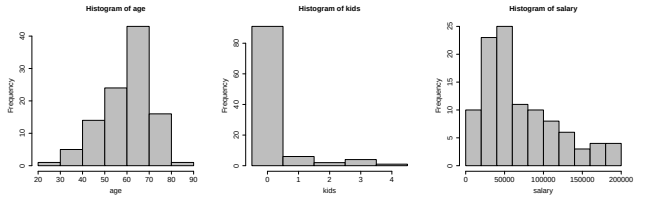
The data

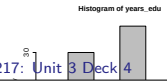
- Feb 21-27, 2010 in Laughlin, Nevada
- anonymous survey of conference attenders
- collected by the Dept of Sociology at Baylor University (Texas)
- 400 surveys distributed, 156 returned, 104 used
- 97 variables
  - UFO beliefs and theories
  - UFO experiences and beliefs about government conspiracies related to UFOs
  - non-UFO paranormal beliefs and experiences
  - religion
  - demographics



The data, $n = 104$

```
> summary(ufo)
      JOBTITLE          age                beliefs    ufo_experience
          :20   Min.   :24.00   cosmic force:60   >=1 :85
Retired   :15   1st Qu.:55.75   god          :44   none:19
RN        : 3   Median :62.00
Teacher   : 2   Mean   :60.62
Therapist: 2   3rd Qu.:68.00
(Other)   :60   Max.   :83.00
NA's      : 2
believe_bigfoot     sex          kids            salary          years_edu
no :17          female:48   Min.   :0.00   Min.   :  6597   Min.   :10.00
yes:87          male  :56   1st Qu.:0.00   1st Qu.: 33365   1st Qu.:14.00
                            Median :0.00   Median : 53983   Median :16.00
                            Mean   :0.25   Mean   : 69757   Mean   :15.97
                            3rd Qu.:0.00   3rd Qu.: 96522   3rd Qu.:18.00
                            Max.   :4.00   Max.   :199258   Max.   :22.00
```





Research question 1

Which null hypothesis?

The average years of education of the general population is 13, whereas the average years of education among the 104 respondents is 16. Is average years of education for those interested in UFOs the same as the general population?

1. $H_0 : \mu_1 = \mu_2$
2. $H_0 : \bar{x}_1 = \bar{x}_2$
3. $H_0 : \mu_0 = 13$
4. $H_0 : \mu_0 = 16$
5. $H_0 : \mu = 16$
6. $H_0 : \mu = 13$
7. $H_0 : \bar{x} = 13$
8. $H_0 : \bar{x} = 16$

Research question 2

Which method?

Is whether or not an individual had a UFO experience associated with years of education?

1. one sample z-test
2. two sample z-test
3. chi-squared test
4. one sample t-test
5. two sample t-test
6. paired t-test
7. ANOVA
8. Linear regression

Research question 3

Which method?

Is there an association between gender and belief in big foot?

1. one sample z-test
2. two sample z-test
3. one sample t-test
4. two sample t-test
5. paired t-test
6. ANOVA
7. Linear regression

Research question 4

Which method?

Is salary associated with years of education?

1. one sample z-test
2. two sample z-test
3. chi-squared test
4. one sample t-test
5. two sample t-test
6. paired t-test
7. ANOVA
8. Linear regression

Research question 5

Which method?

Suppose we categorize years of education as <high school, high school, and >high school, and we want to determine if there is an association between salary and level of education.

1. one sample z-test
2. two sample z-test
3. chi-squared test
4. one sample t-test
5. two sample t-test
6. paired t-test
7. ANOVA
8. Linear regression

Contingency tables

	Beliefs		
	Cosmic Force	God	Total
Bigfoot Yes	51	36	87
Bigfoot No	9	8	17
Total	60	44	104

Which of the following is false?

1. Among those who believe in God, 36/44=81.8% believe in bigfoot.
2. Among those who believe in cosmic force, 51/104=49.0% believe in bigfoot.
3. Overall, the proportion of individuals who believe in bigfoot is higher than the proportion of individuals who believe in God.
4. All of the above are true.

Correlation

The correlation between years of education and salary is 0.22. This means that

1. As annual salary increases by \$1, education increases by 0.22 years.
2. As education increases by one year, annual salary increases by \$0.22.
3. Since the correlation is not 0, we can predict a salary perfectly from years of education.
4. The relationship between salary and years of education follows a curve rather than a straight line.
5. As one of these variables increases, there is a tendency for the other variable to increase also.


```
Call:
lm(formula = salary ~ years_edu)

Residuals:
   Min     1Q Median     3Q    Max
-64379 -31667 -17831  23038 137490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     5024      27855   0.180   0.8572
years_edu       4053       1720   2.356   0.0204 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 47010 on 102 degrees of freedom
Multiple R-squared: 0.05163, Adjusted R-squared: 0.04233
F-statistic: 5.553 on 1 and 102 DF,  p-value: 0.02036
```

What is the estimated regression equation to predict salary by years of education ?

1. $\hat{y} = 5024 + 4053 \times \textit{years\_edu}$
2. $\hat{y} = 5024 + 4053 \times \textit{salary}$
3. $\hat{y} = 4053 + 5024 \times \textit{salary}$
4. $\hat{y} = 4053 + 5024 \times \textit{years\_edu}$
5. $\hat{y} = 27855 + 1720 \times \textit{years\_edu}$
6. $\hat{y} = 1720 + 27855 \times \textit{salary}$

SLR

```
Call:
lm(formula = salary ~ years_edu)

Residuals:
   Min     1Q Median     3Q    Max
-64379 -31667 -17831  23038 137490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     5024      27855   0.180   0.8572
years_edu       4053       1720   2.356   0.0204 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 47010 on 102 degrees of freedom
Multiple R-squared: 0.05163, Adjusted R-squared: 0.04233
F-statistic: 5.553 on 1 and 102 DF,  p-value: 0.02036
```

What is the null hypothesis tested on the line where the p-value is 0.0204?

1. H_0: $\mu_d = 0$
2. H_0: $\beta_0 = 0$
3. H_0: $\beta_1 = 0$
4. H_0: $\mu_1 = \mu_2$

```
> cor(ufo$years_edu,ufo$salary)
[1] 0.2272182
```

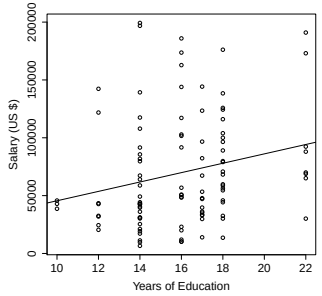
```
Residuals:
   Min     1Q Median     3Q    Max
-64379 -31667 -17831  23038 137490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     5024      27855   0.180   0.8572
years_edu       4053       1720   2.356   0.0204 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 47010 on 102 degrees of freedom
Multiple R-squared: 0.05163, Adjusted R-squared: 0.04233
F-statistic: 5.553 on 1 and 102 DF,  p-value: 0.02036
```

Which of the following statements is true?

1. We have evidence of an association between salary and years of education, and the association is strong.
2. Although we have evidence of an association between salary and years of education, the association is weak.
3. We do not have evidence between salary and years of education.



Which conditions of linear regression are not satisfied?

1. independence of observations
2. linear relationship between x and y
3. constant variability in y about the regression line
4. no conditions are violated

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
source	2	267	133.3	0.477	0.623
Residuals	101	15932	279.5		

How can we interpret these ANOVA results?

1. We have evidence that the mean salary is the same for all three levels of education.
2. We have evidence that the mean salary is different for all three levels of education.
3. We have evidence that the mean salary differs for at least two levels of education.
4. We do not have evidence that the mean salary differs by level of education.


```
 Welch Two Sample t-test

data:  years_edu by ufo_experience
t = -2.0956, df = 25.808, p-value = 0.04608
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
[redacted]
sample estimates:
 mean in group >=1 mean in group none
          15.70588           17.15789
```

What can we say about 95% CI for $\mu_1 - \mu_2$?

1. it would contain $\bar{x}_1 = 15.7$
2. it would contain $p = 0.046$
3. it would contain zero
4. it would not contain zero

Sampling Distributions

For which of the following scenarios is the sampling distribution of the sample mean approximately normally distributed?

1. Population is right skewed and $n = 10$
2. Population is left skewed and $n = 40$
3. Population is normal and $n = 10$
4. 1, 2 and 3
5. 2 and 3 only

Descriptive statistics

- summarizing and visualizing categorical variables
- summarizing and visualizing quantitative variables
- statistics that are/are not robust to outliers

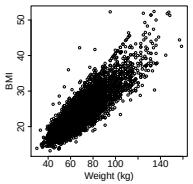
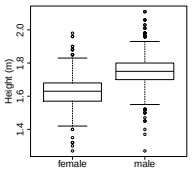
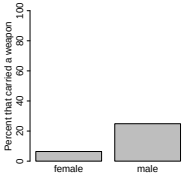
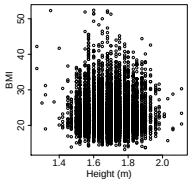
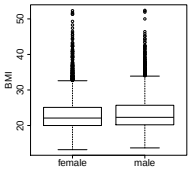
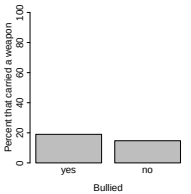
Back

Study design

- recognizing observational vs experimental studies
- potential sources of bias in a study (sampling bias, nonresponse bias, response bias)
- recognizing types of observational study design (simple, stratified, cluster)
- recognizing response vs explanatory variables
- identifying potential confounding variables
- association does not imply causation

Back

Associations

	quantitative-quantitative r, b_1	categorical-quantitative $\bar{x}_1$, $\bar{x}_2$	categorical-categorical $\hat{p}_1$, $\hat{p}_2$
Association?			
No association?			

We need formal statistical tests to determine the direction, magnitude, and significance of the association!

Probability and Distributions

Distributions:

- normal/z, t
- 68 - 95 - 99.7 rule for normal

Interpreting a probability:

- is it large or small?
- would the event be likely to occur by random chance alone?

Back

Sampling Distributions

- For a random sample of size n from a population with proportion p, then when $np \geq 10$ and $n(1-p) \geq 10$ the **sampling distribution** of the **sample proportion** is normally distributed with $\text{mean}(\hat{p}) = p$ and $\text{sd}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$.
- When underlying population distribution is normally distributed or sample size large enough such that CLT applies ($n > 30$), then when sampling from a population with mean μ and standard deviation σ the **sampling distribution** of the **sample mean** is normally distributed with $\text{mean}(\bar{x}) = \mu$ and $\text{sd}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$.

Back

Confidence intervals

- used to *estimate* plausible values for a parameter of interest (like a mean or a proportion)
- CIs take the form

$$\textit{estimate} \pm (z^* \text{or } t^*) \times \textit{se}$$

- true meaning: in the long run, if we took many samples and calculated many confidence intervals, 95% of 95% CIs would actually capture the true parameter value
- know how to interpret
- understand how a confidence interval changes when n, s, or the confidence level changes

Back

Hypothesis tests

1. Define the parameter of interest
2. State the null and alternative hypotheses
3. Identify the appropriate test
4. State assumptions
5. Calculate the test statistic

$$\text{test statistic} = \frac{\text{sample statistic} - \text{null hypothesis value}}{\text{standard error of the sample statistic}}$$

6. Calculate the p-value (for STAT 217, provided by R!)
7. State your conclusion

Back

Types of Errors

Table: Possible outcomes of a hypothesis test

Decision based on observed data	**Unknown Truth**	
	H_0 true	H_0 false
Fail to reject H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

Confidence intervals and hypothesis tests results should agree:

- When you reject H_0, the corresponding CI **should not** include the null value tested in H_0.
- When you fail to reject H_0, the corresponding CI **should** include the null value tested in H_0.

Back

Overview of Statistical Methods

Quantitative variable - means

- one sample t-test
- paired t-test
- two sample t-test
- anova

Categorical variable - proportions

- one sample z-test
- N/A for STAT 217
- two sample z-test
- chi-squared test

Neither a mean or a proportion: simple linear regression.

Back

Different methods

Method	Use	Variables	Estimation	Testing
Single proportion *(one-sample z-test)*	categorical response in single group	one categorical variable	CI for p	H_0: $p = p_0$
*Two proportions *(two-sample z-test)*	categorical response in two groups	two categorical variables	CI for $p_1 - p_2$	H_0: $p_1 = p_2$
*Chi-squared test	categorical response in ≥ 2 groups	two categorical variables	N/A	H_0: no association/ vars independent
Single mean *(one-sample t-test)*	quantitative response in single group	one quantitative variable	CI for μ	H_0: $\mu = \mu_0$
*Two means *(two-sample t-test)*	quantitative response in two groups	one quantitative variable and one categorical variable	CI for $\mu_1 - \mu_2$	H_0: $\mu_1 = \mu_2$
*Dependent means *(paired t-test)*	quantitative response measured on same observation	two paired quantitative variables	CI for μ_d	H_0: $\mu_d = 0$
*ANOVA	quantitative response in > 2 groups	one quantitative variable and one categorical variable	Tukey pairwise intervals	H_0: $\mu_1 = \mu_2 = \cdots = \mu_g$
*Linear regression	quantitative response and a quantitative explanatory	2 quantitative variables	CI for β_1	H_0: $\beta_1 = 0$

*The starred methods can answer the question "Is there an association?" If we reject H_0, then we conclude that some sort of association is present in the two variables.

Back

Test statistics and Confidence Intervals

Method	H_0	Test Statistic	Confidence Interval
Single proportion	$p = p_0$	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Two proportions	$p_1 = p_2$	$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{se}$	$(\hat{p}_1 - \hat{p}_2) \pm z^* \times se$
Chi-squared test	vars independent	$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$	N/A
Single mean	$\mu = \mu_0$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$\bar{x} \pm t^*_{df=n-1} \times s/\sqrt{n}$
Two means	$\mu_1 = \mu_2$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df=given} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Dependent means	$\mu_d = 0$	$t = \frac{\bar{x}_d - 0}{s_d/\sqrt{n}}$	$\bar{x_d} \pm t^*_{df=n-1} \times s_d/\sqrt{n}$
ANOVA	$\mu_1 = \mu_2 = \cdots = \mu_g$	N/A	Tukey
Linear regression	$\beta_1 = 0$	$t = \frac{\hat{\beta}_1 - 0}{se_{\hat{\beta}_1}}$	$\hat{\beta}_1 \pm t^*_{df=n-2} \times se_{\hat{\beta}_1}$

- When performing a statistical analysis with data in R, R by default assumes the two-sided alternative hypotheses as presented above, and all p-values presented represent the final p-values (ie, no need to multiply by two).

Back